

Visualization / Data

- Data Participants:

- A. Choudhary
- T. Critchlow
- L. Diachen
- P. Heermann
- R. Moore
- B. Parvin
- N. Samatova
- A. Shoshani
- M. Vouk

- Vis Participants

- A. Breckenridge
- W. Bethel
- P. Crossno
- J. Kohl
- A. McPherson
- M. Papka
- J. van Rosendale

Petascale Data Implications

- Visualization is a knowledge generation process that requires a human in the loop
- Visualization is a form of qualitative data analysis
- Information extraction(feature detection) is a quantitative process
- Knowledge is the generation of relationships between features
- For petascale data, knowledge generation is essential

Petascale Computers

- Petascale computer may be a heterogeneous environment, with varying data locality
 - View petascale computer as a grid linking heterogeneous resources
 - Requires co-scheduling of CPU, data movement, I/O bandwidth
 - Challenge is maintaining parallelism across a hierarchical data management process

Barriers

- More data than a person can look at
 - Co-scheduling of CPU, storage, and I/O transport
 - Data management / navigation
 - Scalable algorithms for data analysis
- Information extraction
 - Augment human for generation of knowledge
 - On-the-fly data analysis
- Knowledge generation
- Logical data model abstraction drives knowledge abstraction

Dealing with Petascale Data

- Petascale data management
- Post processing of data
- Knowledge extraction and characterization

1. Petascale Data Management

- Distributed data ingestion
 - Problem set up
 - Fast I/O transport for data intensive computing
 - Storage management co-scheduling with CPU
 - Data pre-staging management / scheduling
- Collection based data management - beyond file systems
 - Use of attributes to specify semantic meaning and features
- Knowledge based data management
 - Use of relationships to describe interactions between features in data and semantics of attributes

2. Post Processing of Data

- Not just getting petabytes to the desk top
But also getting information to the desk top
- Data analysis on distributed data
- Levels of abstraction for visualizing data and knowledge
- Integrated compute / analysis / storage
 - Infinite compute with no storage
 - Infinite storage (sensor data)
 - Optimization of storage and compute resource allocation

2. Processing Infrastructure

- Data support
 - Simulation data
 - Sensor data
 - Derived data products
- Agent-based data analysis
 - Mediators for data extraction
- Knowledge generation
 - Definition of feature space
 - Visualization of feature space
 - Differentiation between anomalies and implied knowledge

2. Grid Technologies

- Dealing with heterogeneity
- Data management - Identification of data, navigation through data, organization of data
 - Generation of metadata
 - Semantic relationships between metadata
- Heterogeneous data sources
 - Metadata integration
 - Heterogeneous data models
 - Monitoring, error recovery, fault tolerance

3. Knowledge Generation

- Gain knowledge/insight from peta-scale data
 - Data model specific characterizations that preserve information
 - Characterization of complexity of data model
- Data agents for feature detection
 - Dynamic extraction of knowledge, on-the-fly analysis
- Data provenance for tracking virtual data products
 - Tracking knowledge required to recreate data
 - Virtual data management - reproducibility of data

3. Knowledge Generation

- Generic technology
 - Knowledge generation from petabytes of data
 - Quantification of insight
 - Qualitative organization of insight
 - Characterization of human visualization as a knowledge extraction process
 - Knowledge management

3. Knowledge Generation

- Characterization of problem space as knowledge generation
 - Application of knowledge generation rules to data on-the-fly
 - Analysis of knowledge for worthiness
- Characterization of visualizations as derived data products

Visualization Challenges

- Data analysis guided visualization - on-the-fly feature detection
- Exploration of large data sets without moving data to desk top
- Algorithms for latency management
- Algorithms for tuning presentation to level of resource through level of detail
- 3D and stereo visualization
 - Human interface
- Visualization of knowledge
- Cognitive human interface

Cross-Area Collaborations

- Operating System
 - Grid computing
 - Co-scheduling of CPU, storage, and I/O bandwidth
 - Distributed data management
 - replication vs Internet Backplane Protocol for usage-based data movement
 - Common data format
 - Fault tolerance
 - Bandwidth allocation
 - Global name space for persistent objects
 - Storage abstraction for manipulating data within storage

Cross Area Collaboration

- Portability / tools
 - Bandwidth - parallel I/O support from application
 - Grid computing
 - Persistent objects
 - Common data format
- Performance
 - Bandwidth end-to-end
 - Data bottleneck identification
 - Network protocol support for QOS
- Programming models and run time
 - Visualization based simulation problem set-up
 - High-level data abstraction
 - Persistent parallel programming model

Coordination across Communities

- Unification - Interoperability of technologies
- Unification of islands of expertise
 - Unification of approaches between data and vis
- Unification of application-drivers with computer science research

Current Gaps in MICS CS Research

- Application of grid computing capabilities throughout MICS program
 - Intelligent data processing - Data intensive architectures
 - Computation in data resource
 - Agent based systems for data access
 - Fault tolerant computing
- Integration of knowledge generation across data analysis and visualization communities
 - Knowledge extraction and visualization
 - Abstraction of complex data types
- Productization of software
 - porting to infrastructure independent representations (OMG-MDA)